

What is claimed is:

1. A method for controlling performance of an operation in relation to a set of resources within a data processing network, comprising the steps of:

5 computing a set of hash values representing a set of resources for which an operation has been performed;

storing the set of hash values;

in response to a requirement for performance of the operation, computing a new set of hash values representing the set of resources;

10 comparing the new hash values with the stored hash values for the set of resources to identify matches between new hash values and stored hash values;

determining that performance of the operation is not currently required for resources for which a match is identified between the respective new hash value and a stored hash value; and

15 performing the operation for resources for which no match is identified between the new hash value and any stored hash value.

2. The method of claim 1, wherein the operation comprises scanning the resources to identify computer viruses.

20

3. The method of claim 1, wherein the operation comprises making a backup copy of the resources.

25

4. The method of claim 1, for controlling performance of virus scanning and backup copy operations in relation to a set of resources within a data processing network, the method comprising:

30 using said identification of a match between a respective new hash value and a stored hash value for a resource, resulting from a single comparison of new and stored hash values, to determine that neither virus scanning nor backup copy operations are currently required for the resource.

5. The method of claim 1, wherein the step of computing a new set of hash values comprises reading the set of resources from a first storage medium into a second storage medium which provides faster access than the first storage medium and computing the set of hash values, and wherein the method further comprises:

5 comparing each of the set of resources with a maximum size limit to identify resources within said set which are smaller than said size limit, and

retaining said smaller resources within said second storage medium to enable further operations.

10 6. The method of claim 1, wherein the operation comprises transferring a resource across a low bandwidth communication channel.

7. The method of claim 1, wherein the steps of computing hash values comprise:

15 applying a secure hash function to a bit pattern representing a resource, for each of a set of resources.

8. The method of claim 7, wherein the set of resources for which hash values are computed for a data processing system comprises the set of all files of executable file types on the system.

20 9. The method of claim 1, wherein the set of resources are distributed across a plurality of data processing systems within the network and the steps of storing the set of hash values and comparing the new hash values with the stored hash values are performed at a first data processing system within the network for the set of resources distributed 25 across the plurality of data processing systems.

30 10. The method of claim 9, wherein the steps of computing hash values for a resource are performed at a respective one of the plurality of data processing systems at which the resource is stored, the method further comprising sending the computed hash values to said first data processing system.

11. The method of claim 9, further comprising the step of sending, to each data processing system storing a resource for which it is determined that performance of the operation is not currently required, an indication that performance of the operation is not
5 currently required for the resource.
12. The method of claim 9, wherein the step of performing the operation is performed at the first data processing system and the result of performing the operation is communicated to each of the plurality of data processing systems storing a resource for
10 which the operation is required.
13. The method of claim 1, wherein at least one resource of the set of resources comprises a group of files and the step of computing a set of hash values comprises computing a single hash value for the group of files.
15
14. The method of claim 13, wherein at least one resource of the set of resources is a compressed group of files.
15. A method for controlling scanning for computer viruses within a data processing network, comprising the steps of:
20 computing a set of hash values representing a set of resources which have been determined to be virus-free;
25 storing the set of hash values;
representing the set of resources;
comparing the new hash values with the stored hash values for the set of resources to identify matches between new hash values and stored hash values;
determining that no virus scan is currently required for resources for which a match is identified between the new hash value and a stored hash value; and
30 performing a virus scan for each resource for which no match is identified

between the new hash value and any stored hash value.

16. The method of claim 15, wherein the set of resources are distributed across a plurality of data processing systems within the network, and wherein the steps of storing

5 the set of hash values and comparing the new hash values with the stored hash values are performed at a first data processing system within the network for the set of resources.

17. The method of claim 16, wherein the steps of computing hash values for a resource are performed at a respective one of the plurality of data processing systems at 10 which the resource is stored, the method further comprising sending the computed hash values to the first data processing system.

18. The method of claim 16, further comprising the step of sending, to each data processing system storing a resource for which it is determined that a virus scan is not 15 currently required, an indication that the virus scan is not currently required for the resource.

19. The method of claim 16, wherein the step of performing the virus scan is performed at the first data processing system and the result of the virus scan is 20 communicated to each of the plurality of data processing systems storing a resource for which a virus scan is required.

20. The method of claim 16, further comprising the steps of:

in response to determining that a virus scan is currently required for a resource,

25 sending a request for a copy of the resource from the first data processing system to a respective one of the plurality of data processing systems at which the resource is stored,

receiving the copy of the resource at the first data processing system,

performing a virus scan of said resource at the first data processing system, and

reporting the result of the virus scan to the respective one of the plurality of data

30 processing systems at which the resource is stored.

21. The method of claim 15, wherein the step of determining that no virus scan is currently required for resources for which a match is identified comprises the following steps:

5 identifying a subset of resources within said set of resources for which a match is identified between the new hash value and a stored hash value;

determining whether each resource within said subset has been classified virus-free by each of a plurality of virus scans of the resource at scan times which differ by more than a threshold; and

10 determining that no virus scan is currently required for each resource for which a positive determination is made that the resource has been classified virus-free by each of a plurality of virus scans at scan times which differ by more than the threshold.

22. The method of claim 15, further comprising:

15 comparing the new hash value computed for a first resource with hash values computed for other resources of said set of resources, to identify matching hash values indicating replicated resources;

in response to determining that a virus scan is currently required for the first resource, performing a virus scan of the first resource; and

20 recording a result of the virus scan of the first resource as a virus scan result for any replica resources.

23. The method of claim 22, wherein the set of resources is distributed across a plurality of data processing systems within a network, the method further comprising:

25 forwarding an indication of the result of the virus scan to the data processing systems storing the first resource and any replica of the first resource within the network.

24. A method for controlling scanning for computer viruses within a data processing network, comprising the steps of:

30 for a set of resources determined to be virus free, storing a set of hash values

derived from the set of resources;

in response to a subsequent requirement for a virus check, comparing a new set of hash values derived from the set of resources with the stored set of hash values to identify matches between new and stored hash values;

5 identifying the resources for which the new hash values match stored hash values, and determining that a virus scan is not currently required for the identified resources for which the new hash values match the stored hash values; and

initiating a virus scan for resources for which the new hash values do not match stored hash values.

10

25. The method of claim 24, wherein the set of resources are distributed across a plurality of data processing systems within the network, and wherein the steps of storing the set of hash values, comparing the new hash values with the stored hash values, and identifying the resources for which the new hash values match stored hash values, are

15 performed at a first data processing system within the network for the distributed set of resources.

26. The method of claim 25, further comprising the step of sending, to each data processing system storing a resource for which it is determined that a virus scan is not 20 currently required, an indication that the virus scan is not currently required for the resource.

27. The method of claim 25, further comprising the step of performing the virus scan at the first data processing system and communicating a result of the virus scan to each of 25 the plurality of data processing systems storing a resource for which the virus scan is performed.

28. The method of claim 25, wherein the step of initiating a virus scan comprises the steps of:

30 sending a request for a copy of the resource from the first data processing system

to a respective one of the set of data processing systems at which the resource is stored, receiving the copy of the resource at the first data processing system, performing a virus scan of said resource at the first data processing system, and reporting the result of the virus scan to the respective one of the set of data processing systems at which the resource is stored.

5 29. The method of claim 24, wherein the step of determining that a virus scan is not currently required for resources for which a match is identified comprises the following steps:

10 identifying a subset of resources within said set of resources for which a match is identified between the new hash value and a stored hash value;

determining whether each resource within said subset has been classified virus-free by each of a plurality of virus scans of the resource at scan times which differ by more than a threshold; and

15 determining that no virus scan is currently required for each resource for which a positive determination is made that the resource has been classified virus-free by each of a plurality of virus scans at scan times which differ by more than the threshold.

30. A data processing apparatus comprising:

20 a data processing unit;

a data storage unit; and

a repository manager configured to store, in at least one repository within the data storage unit, a set of hash values derived from a set of resources determined to be virus free; and

25 a virus scan coordinator for comparing a new set of hash values derived from the set of resources with the stored set of hash values to identify matches between the new hash values and stored hash values, for identifying resources for which the respective new hash values match stored hash values, for initiating a virus scan for resources for which the respective new hash values do not match stored hash values and for controlling the 30 repository manager to store in the repository an indication that a virus scan is not

currently required for at least some of the identified resources.

31. A data processing apparatus comprising:

a data processing unit;

5 a data storage unit; and

a repository manager configured to store a set of hash values in at least one repository within the data storage unit, wherein the hash values have been derived from a set of resources for which the operation has been performed; and

10 a coordinator for coordinating performance of an operation, for comparing a new set of hash values derived from the set of resources with the stored set of hash values to

identify matches between the new hash values and stored hash values, for identifying resources for which the respective new hash values match stored hash values, and for controlling the repository manager to record in the repository an indication that at least some of the identified resources do not currently require performance of the operation.

15

32. A distributed data processing system comprising:

a first data processing apparatus comprising a data processing unit; a data storage unit; a repository manager configured to store a set of hash values derived from a set of resources in at least one repository within the data storage unit; and a virus scan

20

coordinator for comparing a new set of hash values derived from the set of resources with the stored set of hash values to identify matches between the new hash values and stored hash values, for identifying resources for which the respective new hash values match stored hash values, and for controlling the repository manager to store in the repository an indication that at least some of the identified resources do not currently require a virus

25

scan; and

at least a second data processing apparatus comprising a data processing unit; a data storage unit for storing at least one resource of the set of resources; and a hash value generator, the hash value generator being configured to compute a set of hash values for said at least one resource and to send the set of hash values to the coordinator.

30

33. A distributed data processing network comprising:

at least a first data processing system comprising a data processing unit; a data storage unit; a repository manager configured to store a set of hash values derived from a set of resources in at least one repository within the data storage unit; and a coordinator

5 for coordinating performance of an operation, for comparing a new set of hash values derived from the set of resources with the stored set of hash values to identify matches between the new hash values and stored hash values, for identifying resources for which the respective new hash values match stored hash values, and for controlling the repository manager to record in the repository an indication that at least some of the

10 identified resources do not currently require performance of the operation; and

at least a second data processing system comprising a data processing unit; a data storage unit for storing at least one resource of the set of resources; and a hash value generator, the hash value generator being configured to compute a set of hash values for said at least one resource and to send the set of hash values to the coordinator.

15 34. The distributed data processing network of claim 33, wherein the coordinator of the first data processing system is configured to coordinate performance of the operation in relation to a set of resources distributed across a plurality of data processing systems within the network, and the steps of storing the set of hash values and comparing the new hash values with the stored hash values are performed at the first data processing system within the network for the set of resources distributed across the plurality of data processing systems.

20 35. The distributed data processing network of claim 34, further comprising a plurality of data processing systems each comprising a coordinator for coordinating performance of the operation in respect of resources distributed across a respective plurality of data processing systems.

25 36. A computer program product, comprising program code recorded on a recording medium, for controlling the performance of operations on a data processing system on

which the program code executes, wherein the program code comprises:

 a repository manager configured to store a set of hash values in at least one repository, for a set of resources determined to be virus free; and

 a virus scan coordinator for comparing a new set of hash values derived from the 5 set of resources with the stored set of hash values to identify matches between the new hash values and stored hash values, for identifying resources for which the respective new hash values match stored hash values, and for controlling the repository manager to store in the repository an indication that at least some of the identified resources do not currently require a virus scan.

10

37. A method for controlling scanning for computer viruses within a data processing network, comprising the steps of:

 installing a set of virus-free decoy resources on a data processing system;

 storing a set of hash values derived from the set of decoy resources;

15 in response to a subsequent requirement for a virus check, comparing a new set of hash values derived from the set of decoy resources with the stored set of hash values to identify matches between new and stored hash values;

 classifying decoy resources for which the new hash values match stored hash values as virus-free; and

20 classifying decoy resources for which the new hash values do not match any stored hash values as contaminated.

38. The method of claim 37, further comprising:

 installing virus-free decoy resources on a plurality of data processing systems 25 within the network; and

 in response to the classification of decoy resources as contaminated, generating a report of the distribution of contaminated resources across the plurality of data processing systems.